



2176
#5

THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants: D. Meyerzon et al.

Attorney Docket No. MSFT112958

Application No.: 09/493,748

Group Art Unit: 2776

Filed: January 28, 2000

Examiner: --

Title: ADAPTIVE WEB CRAWLING USING A STATISTICAL MODEL

COMMUNICATION REGARDING TRANSLATION

RECEIVED

Seattle, Washington 98101

OCT 07 2003

September 29, 2003

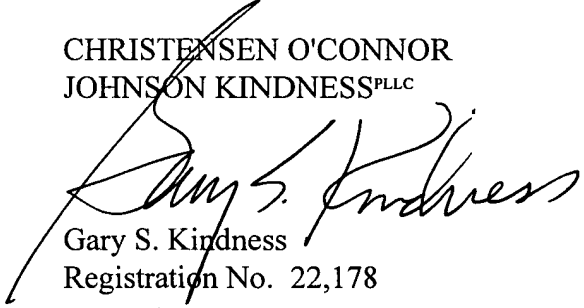
Technology Center 2100

TO THE COMMISSIONER FOR PATENTS:

Attached is a translation of Japanese Patent Document No. 11-328191 cited in an Information Disclosure Statement previously filed in the above-identified application. While applicants have no reason to doubt that the translation is accurate since it was obtained from a commercial source, applicants make no representations in this regard.

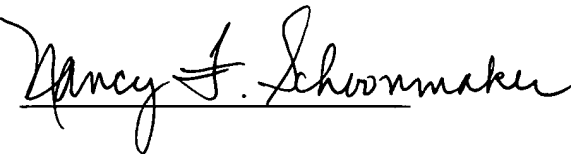
Respectfully submitted,

CHRISTENSEN O'CONNOR
JOHNSON KINDNESS^{PLLC}


Gary S. Kindness
Registration No. 22,178
Direct Dial No. 206.695.1702

I hereby certify that this correspondence is being deposited with the U.S. Postal Service in a sealed envelope as first class mail with postage thereon fully prepaid and addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the below date.

Date: 9/29/03



GSK:pbe/nfs

LAW OFFICES OF
CHRISTENSEN O'CONNOR JOHNSON KINDNESS^{PLLC}
1420 Fifth Avenue
Suite 2800
Seattle, Washington 98101
206.682.8100

[Translation]



(19) Japan Patent Office (JP)
(12) **PATENT RELEASE REPORT (A)**

(11) Patent Application Release No.
Patent Release **Hei. 11-328191**

(43) Release date: November 30, 1999

(51) Int.Cl. ⁵	Identification Symbol	F1	
G 06 F 17/30		G 06 F 15/40	310F
			380Z
		15/403	340B

Examination requested: Yes
Items in Application: 6 OL (Total 7 pages)

(21) Application No.: Patent Application 10(1998)-129829

(22) Application date: May 13, 1998

RECEIVED

(71) Applicant: 000004237
NEC Corp., Ltd.
7-1 Shiba 5-chome, Minato-ku **Technology Center 2100**
Tokyo [Japan]

OCT 07 2003

(72) Inventor: Takashi Kato
c/o NEC Corp., Ltd.
[same address]

(74) Agent: Katsumi Udaka, Patent attorney

(54) **Name of invention:** WWW Robotic Retrieval System

(57) **Summary** (Amendment exists)

Issue To provide a method for automatically deriving retrieval standards not for simple retrieval, but for anticipating already updated WWW pages and retrieving only said pages.

Means of resolution With updating-frequency operating engine 22 to operate the updating frequency of any WWW page from the retrieved results, updating-expectation operating engine 23 to run the updating expected values at any time from the updating frequency and engine 24 to create

retrieval-sequence tables that automatically extract retrieval priority rankings, it anticipates whether-at a particular moment, any WWW page has been updated from updating-frequency operating engine 32 and updating-expectation operating engine 33 and then makes up the retrieval-order table. The WWW robotic retrieval system retrieves according to this retrieval-order table.

[Diagram on page 1 is identical to Figure 1 and so is not translated here. -- Translator]

Scope of Patent Application

Application Item 1 A WWW robotic retrieval system characterized by doing retrieval according to a retrieval order automatically created when WWW pages are retrieved.

Application Item 2 A WWW robotic retrieval system characterized by specifying appropriate reference points from the retrieval reference points of line numbers that can be starting points when doing a WWW page retrieval, and by doing a retrieval according to an automatically created retrieval sequence.

Application Item 3 A WWW robotic retrieval system characterized by automatically extracting any WWW page's updating frequency based on the structure of the WWW page found by the WWW robot.

Application Item 4 A WWW robotic retrieval system characterized by automatically extracting the updating expectation at any time from the updating frequency sought in one chapter [?-blurred text] from each WWW page by hyperlink relationships between WWW pages.

Application Item 5 A WWW robotic retrieval system characterized by automatically creating the priority sequencing for the retrieval objective when the WWW robot retrieves according to each WWW page's updating frequency and retrieval expectations.

Application Item 6 A WWW robotic retrieval system characterized by being equipped with -

- An updating-frequency calculating engine to calculate the updating frequency of each WWW page,

- An updating expectation calculating engine to calculate the updating expectation at any time from each WWW page's updating frequency and
- An engine to create retrieval sequence tables by automatically extracting the retrieval sequencing priority when the WWW robot retrieves from updating frequency values and updating expectation values.

Detailed Explanation of Invention

0001 Technical Field to Which Invention Belongs This invention relates to methods of extracting retrieval reference points when retrieving with a WWW (Worldwide Web) robot, and specifically bears on a system for WWW robotic search system for finding the optimal retrieval sequence from multiple search reference points that constitute starting points when doing retrievals with WWW robots.

0002 Usual Technology WWW robotic retrieval systems are characterized by having the capacity to detect WWW page structures within WWW servers and to detect WWW page updating. WWW robotic search systems do searches of WWW servers for a specific WWW server's top page or for the starting point of a specific WWW page, analyze—from the WWW page obtained in the retrieval—the HTML (Hyper Text Makeup Language) describing the WWW page's information and retrieve the location of the next WWW page hypered from this WWW page. HTML is a language for describing hypertext according to a HTTP (Hyper Text Transfer Protocol), which is a protocol for clients and servers to communicate on the WWW.

0003 Also, this search system records data on the contents and location of newly created, altered or deleted WWW pages found in the course of searching WWW servers. WWW robots extract the WWW pages sought on WWW servers by the above-described sequences.

0004 Still, in past WWW-search robots operators have always specified in advance the search-start locations as well as search sequencing. WWW-search robots' searching process proceeds from the start points and sequences pre-set by an operator and proceed on the basis of rules sequentially analyzing the HTML to acquire WWW pages.

0005 In such systems using, as is, the parameters pre-set by an operator, when accessing many WWW pages over a wide

range, the search time, range over which the WWW robot searches and load on the network are [word blurred] proportionate to the number of WWW servers being searched and the depth of the Hyperlink.

0006 In the usual document search system, when retrieving documents containing words or phrases designated by an operator, document retrieval systems are proposed that are equipped with a function to retrieve from all documents, are equipped with key-word retrieval functions to reference indexes made up of previously extracted words and phrases from all documents and are equipped with a way to judge— from specified words and phrases and other conditions— whether both of these function should be used and deciding, based on that, to do the more useful of the retrievals. (E.g., Patent Hei.10[1998]-21255] However, this document retrieval system is one that decides whether whole-text retrieval or a key-word retrieval is advantageous and is conducted according to reference points at the start of the retrieval and specified reference points.

0007 In these usual WWW robotic retrieval systems WWW pages are always continuously retrieved that are hyperlinked sequentially from starting points an operator has designated. But, the structure of WWW pages in current WWW servers is complex and so quite deep even in the hyperlink stage. So, under the usual retrieval rules by which an operator gives starting points in advance and sequential retrievals are done based on that data, the problem is noted that a great amount of time is needed to get to WWW pages actually altered within WWW servers. And, from the fact that WWW server retrievals are done via networks, there is the problem network resources being wasted in doing much unnecessary retrieval, increasing the load on the networks.

0008 Issues the Invention Seeks to Resolve The theme of this invention is to provide a WWW robotic retrieval system that will solve the usual technical problems as described above, educe the updating frequencies and updating expectations of each WWW page in WWW servers and from these automatically educing the retrieval sequences of optimal update starting points.

0009 Means to Resolve the Issues This invention conducts retrievals of actual WWW pages by automatically deciding on priority ordering for WWW robotic retrieval systems that

search out WWW pages. More concretely, as shown in Figure 1, the WWW robotic retrieval system is equipped with -

- Updating-frequency calculating engine 22 to calculate the updating frequency of each WWW page,
- Updating expectation calculating engine 23 to calculate the updating expectation at any time from each WWW page's updating frequency and
- Engine 24 to create retrieval sequence tables by automatically extracting the retrieval sequencing priority when the WWW robot retrieves from updating frequency values and updating expectation values.

0010 Effects WWW server page structural memory 32 records structural data showing information about WWW pages and connections between those pages in WWW servers that result from retrievals done by WWW server retrieval engine 22.

WWW-server page updating frequency operating engine 22 uses WWW-page structural data to calculate the value of WWW page updating frequencies and stores those results in WWW-server page-updating frequency memory 33.

0011 WWW-server page updating expectation engine 23 automatically creates parameters for computing expectations for each WWW page being updated at a certain time from data in updating frequency memory 33 and records these parameters in WWW server-page updating expectation memory 34. Engine 24 for creating the WWW server sequencing tables automatically creates tables that give sequencing for which WWW server's page is to be retrieved when WWW server retrieval engine 21 is conducting a retrieval.

0012 Format for Effectuating the Invention Next, I will give a detailed explanation of the format for effectuating this invention, referring to the figures. In Figure 1, this invention's first format for implementing this invention includes -

- Input/output device 1 such as the keyboard and display,
- Computing processing section 2 that runs by program controls,
- Memory device 3 which holds data, and
- Network communication device 4, which exchanges data with external WWW servers via the Network, etc.

0013 Memory device 3 is equipped with retrieval schedule 31, WWW server page-structure memory 32, WWW server page

updating frequency memory 33, WWW server page updating expectation memory 34 and retrieval sequencing table 35.

0014 Retrieval schedule memory 31 stores [blurred word] data on times required in the past for WWW servers' starting times for retrievals and time spent on retrievals.

0015 WWW server page-structure memory 32 stores WWW page text contents arrayed within WWW servers and the kind of connecting relationship each of these WWW pages has to another.

0016 WWW server page updating frequency memory 33 calculates the updating frequency rate of each WWW page and stores this as numeric data.

0017 WWW server page updating expectation memory 34 calculates the possibility that each WWW page may be expected to be updated at any point in time and stores this as numeric data.

0018 When one is doing the next retrieval, the retrieval sequence table 35 holds data giving the priority order from the standpoint of which WWW page is to be searched from which WWW server.

0019 Operating control 2 is equipped with -

- WWW server search engine 21,
- WWW server page-update frequency operating engine 22,
- WWW server page-update expectation operating engine 23 and
- Engine 24 for creating a WWW server search sequence table.

0020 Using a run command from input/output device 1, WWW server search engine 21 carries out external WWW server page retrievals via network device 4. Such data as retrieval results, WWW page updating and deletions of additions of new WWW pages and hypering relationships between WWW pages are stored in WWW-page structural memory 32. Data on times that WWW pages were retrieved are stored in retrieval schedule memory 31.

0021 WWW server page updating expectation operating engine 23 newly calculates the updating frequency of each WWW page based on (1) WWW page updating data and hypering data that WWW server page structure memory 32 has recorded and on

(2) data relating to the updating frequency of each WWW page which WWW server page updating frequency memory 33 has recorded; and it stores the calculated results in WWW server page updating frequency memory 33.

0022 Based on data that WWW server page updating expectation memory 33 has recorded, WWW server page update expectation calculator engine 33 calculates the degree that updating is to be expected for each WWW page at any point in time; and it records those results in WWW server page update expectation memory 34.

0023 Engine 24 for creating a table of WWW server retrieval sequencing calculates retrieval sequencing from keys input from input/output device 1 and from data that WWW server page updating frequency memory 33 and WWW server page updating expectation memory 34 have recorded; and it stores those results in retrieval sequencing table 35.

0024 Explanation of Operation Next, I will explain in detail how the makeup from this invention operates, referring to Figures 1, 2, 3, 4 and 5.

0025 Upon receiving a run command signal from input/output device 1, WWW server retrieval engine 21 uses network communication device 4 to acquire WWW page data from an outside WWW server according to retrieval sequencing that retrieval sequencing table 35 has recorded in advance. WWW page data acquired by the retrieval is of four kinds:

- "HTML describing WWW page is just retrieved,"
- "Has WWW page just retrieved been updated since its previous retrieval?"
- "WWW page that is hypered original of WWW page just retrieved," or
- "WWW page that is hypered derivative of WWW page just retrieved."

WWW server retrieval engine 21 stores data on these four WWW page types in WWW server page structure memory 32.

0026 Based on data that WWW server page structure memory 32 has recorded, WWW server page update frequency calculating engine 22 calculates and determines the WWW server page updating frequency to show the frequency of WWW page updating.

0027 The WWW page structure in a WWW server as retrieved by WWW server retrieval engine 21 has a form equivalent to

a neural net of the perceptron [from Japanese phonetics. Translator] type as in Figure 2. So, with this invention, treating each WWW page as a neural net node and considering the weight that updating frequency has for each WWW page (node) as corresponding to the updating a WWW page's updating frequency, one calculates each WWW page's updating frequency. When a WWW page has the structure as in Figure 2, WWW server page updating frequency-calculating engine 22 uses equation (1) shown in Figure 2 to find any WWW page_a's updating frequency from the updating frequency of the derivative WWW page hypered from WWW page_a.

0028 As its method for drawing out the updating frequency of all WWW pages in a WWW server, WWW server page updating frequency calculating engine 22 picks out a WWW page at random and calculates the updating frequency of that selected WWW page. I will now explain the sequence for deciding on the WWW page to select.

0029 First, WWW server page updating frequency calculating engine 22 looks at the derivative WWW page located at the lowest position on the WWW page that WWW server page structure memory 32 has recorded and extracts this WWW server page's latest frequency. Second, it marks the page for which it has just extracted the updating frequency. Third, WWW server page updating frequency calculating engine 22 makes the derivative WWW page located at the lowest place on the WWW page without a mark and which WWW server page structure memory 32 has recorded the next WWW page to be looked at. Thereafter, by repeating this sequence, it obtains the WWW pages' updating frequency.

0030 When eight WWW pages have been hypered in a relationship like that shown in Figure 3, the calculating sequence can be calculated, for instance, in the order of "page 8 → page 5 → page 6 → page 7 → page 2 → page 3 → page 4 → page 1" as shown in the box below Figure 3.

0031 WWW server page updating frequency calculating engine 22 works out each WWW page's updating frequency by the method discussed above, and that is recorded in WWW server page updating frequency memory 33.

0032 WWW server page updating expectation engine 23 calculates a value for the degree to which at a certain point in time a WWW page may be updated, and that result is recorded in WWW server page updating expectation memory 34.

0033 WWW server page updating expectation engine 23 calculates the degree of expectation for each WWW page's updating by the following method. When calculating the expectation of a WWW page being updated, WWW page updating expectation calculation engine 23 uses [a] the time elapsed since the previous retrieval was carried out, recorded in retrieval schedule memory 31, up to the time when the last retrieval was done and [b] updating frequency w which is recorded in WWW server page updating frequency memory 33 and uses Figure 4's equation (2) to calculate parameter ε in Figure 4's equation (1) to obtain the updating expectation degree Ex . The degree of updating expectation Ex at any moment in time t is found with Figure 4's equation (1).

0034 The sequence by which WWW server retrieval order table creating engine 24 compiles the WWW retrieval robot's retrieval sequencing data is as follows. Engine 24 first copies to retrieval sequence table 35 the WWW page data recorded by WWW server page structure memory 32 (Table 5-1 of Figure 5).

0035 Second, engine 24 calculates the degree of confidence of each WWW page's being updated, using data from WWW server page updating expectation memory 34 and from retrieval schedule memory 31; and it records the results in retrieval sequencing table 35. (Figure 5's Table 5-2)

0036 Third, engine 24 treats the higher value WWW page updating expectation numbers as as the higher priority data of retrieval sequencing table 35 and realigns them into a priority order with high [blurred word]. (Figure 5's Table 5-3)

0037 Fourth, following conditions that people (operators) have given to input/output device 1, engine 24 deletes from the data recorded in retrieval sequence table 35 those WWW page data not matching those conditions. (Figure 5's Table 5-4)

0038 Effectiveness of Invention The results obtained by the WWW robotic retrieval system of this invention create automatically the degree of retrieval expectation at a certain time for a WWW page, using the retrieval conditions of the subject WWW page and the derivative WWW page which that page has hypered; and based on that result it decides the priority sequencing for the next retrieval time. Thus, the WWW robot will proceed to retrieve first WWW pages with

a high degree of expectation of their being updated. That makes it possible to quickly acquire WWW page data that has been updated.

0039 Also, following conditions that people (operators) have given the retrieval table when making it up, one can weed out WWW pages with no retrieval value (Fig. 5 Table 5-4), so that unnecessary retrieval operations are reduced. That makes it possible to cut back the load on networks generated when a WWW robot retrieval system does retrievals.

Simple Explanation of Figures

Figure 1 is a block diagram showing the makeup of the format for effectuating this invention.

Figure 2 is a diagram illustrating the method for calculating the updating-frequency operation of the format for effectuating this invention.

Figure 3 is a diagram illustrating the method of calculating the updating frequency operation of the format for effectuating this invention.

Figure 4 is a diagram illustrating the method of calculating the updating expectation of the operation of the format for effectuating this invention.

Figure 5 is a diagram illustrating the method for creating the retrieval table in the operation of the format for effectuating this invention.

Keying Symbols

- 1 Input/output device
- 2 Calculation processor
- 3 Memory device
- 4 Network communication device
- 21 WWW server retrieval engine
- 22 WWW server page-update frequency calculator engine
- 23 WWW server page-update expectation calculator engine
- 24 WWW server update retrieval sequence table-creating engine
- 31 Retrieval schedule memory
- 32 WWW server page structure memory
- 33 WWW server page update frequency memory

34 WWW server page update expectation memory
 35 WWW server update sequencing table

[Explanation of Japanese text in figures -- Translator]

Figure 1: [See keying symbols above]

Figure 2: Format of current page

i₁ - Derivative page 1
 i₂ - Derivative page 2
 i₃ - Derivative page 3

Figure 3: [Each circle is a numbered "page."]

Sequence for doing calculation is as follows:
 [Box with numbered "pages"]

Figure 4:

[vertical axis] Updating Expectation Value Ex

Expectation value Ex

[horizontal axis] Time t

At time t_u the expectation value Ex_u of the WWW page being updated is found with the equation below: [Subscript letters are too blurred to read; assumed to be "u" for update.]

$$Ex_u = \frac{1}{1 + \exp(\epsilon_u + t_u)} \quad \dots (1)$$

ε_u is found using the equation below from the average value T_[subscripts unclear] of the [blurred word] time integrals of WWW server retrievals:

$$E_u = T_{[unclear \text{ subscript letter}]} + \log \left(\frac{1}{w_u} - 1 \right) \quad \dots (2)$$

Figure 5:

Table 5-1 Page No. Updating expectation

[pages 1-8]

[Text over arrow between Tables 5-1 and 5-2:] Calculates update frequency of each [illegible] page.

[Same column headings in 5-2, 5-3 and 5-4 as in 5-1]

[Text beside arrow between 5-2 and 5-3] Lined up in order of higher updating frequency.

[Text over arrow between 5-3 and 5-4] Following conditions provided from input/output device, unneeded pages are deleted from list.

[Text over arrow between 5-3 and 5-4] (Note) Condition: Retrieves only those with degree of expectation over 0.5.